Weekly Formative Exams and Creative Grading Enhance Student Learning in an Introductory Biology Course

E. G. Bailey,[†]* J. Jensen,[†] J. Nelson,[‡] H. K. Wiberg,[§] and J. D. Bell[‡‖]
[†]Department of Biology, [‡]Department of Physiology and Developmental Biology, and [§]Department of Statistics, Brigham Young University, Provo, UT 84602

ABSTRACT

First-year students often become discouraged during introductory biology courses when repeated attempts to understand concepts nevertheless result in poor test scores. This challenge is exacerbated by traditional course structures that impose premature judgments on students' achievements. Repeated testing has been shown to benefit student ability to recognize and recall information, but an effective means to similarly facilitate skill with higher-order problems in introductory courses is needed. Here, we show that an innovative format that uses a creative grading scheme together with weekly formative midterm exams produced significant gains in student success with difficult items requiring analysis and interpretation. This format is designed to promote tenacity and avoid discouragement by providing multiple opportunities to attempt demanding problems on exams, detailed immediate feedback, and strong incentives to retain hope and improve. Analysis of individual performance trajectories with heat maps reveals the diversity of learning patterns and provides rational means for advising students.

INTRODUCTION

Although the need to design science courses that emphasize development of cognitive skills beyond information acquisition is well recognized, implementation of successful approaches remains a challenge, especially for first-year students (Seymour and Hewitt, 1997; American Association for the Advancement of Science, 2011; President's Council of Advisors on Science and Technology, 2012). A promising means of achieving that implementation may be to take advantage of the "testing effect" by administering frequent cumulative exams (Balch, 1998; Roediger and Karpicke, 2006a,b; Marsh *et al*., 2007; Wickline and Spektor, 2011). The testing effect is evidenced by enhanced retrieval of target information that was previously tested compared with simply rereading or restudying the material. Although the majority of work on the testing effect has focused on low-level memory tasks (Carrier and Pashler, 1992; Carpenter and DeLosh, 2006; Carpenter and Pashler, 2007; Carpenter *et al*., 2008, 2009; Chan and McDermott, 2007; McDaniel *et al*., 2007; Johnson and Mayer, 2009; Rohrer *et al*., 2010), recent research in laboratory settings has suggested that these benefits could extend to items at higher levels of Bloom's taxonomy (Anderson and Krathwohl, 2001; Kang *et al*., 2011; Jensen *et al*., 2014). Recent classroom results also suggest that testing students using items at the application level or above increases student performance on future test questions requiring higher-order thinking skills, even when question content has not been tested before (McDaniel *et al*., 2013; Jensen *et al*., 2014). In addition, we previously achieved a 50% gain in student ability to solve difficult items involving interpretation of experimental data in an upper-division cell biology course by using multiple formative exams (Kitchen *et al*., 2003). Thus, we predicted that repeatedly testing first-year biology students on the same types of higher-order test items throughout the semester would improve their performance.

David Marcey, *Monitoring Editor*
Submitted February 22, 2016; Revised August 30, 2016; Accepted October 3, 2016
CBE Life Sci Educ March 1, 2017 16:ar2
DOI:10.1187/cbe.16-02-0104

[‖]Present address: Brigham Young University–Hawaii, 55-220 Kulanui Street, Building 5, Laie, HI 96762-1293.
*Address correspondence to: E. G. Bailey (liz_bailey@byuh.edu).

Supplemental Material can be found at:
http://www.lifescied.org/content/suppl/2017/01/22/16.1.ar2.DC1.html

While the testing effect is our primary theoretical rationale behind the implementation of these cumulative exams, research suggests that the frequency of testing may also have benefits. In their meta-analysis, Bangertdrowns *et al*. (1991) found an overall positive effect of increasing the frequency of testing on student performance, but the amount of improvement in achievement diminished as frequency increased. However, Leeming (2002) compared an exam-a-day procedure with giving four unit exams and found that the exam-a-day format increased overall performance in the class and performance on a retention test. Sedki (2011) looked at student preference in college and found that students prefer more frequent testing on smaller amounts of material. Phelps analyzed several hundred studies from 1910 to 2010 on the effects of testing on student achievement and found that more frequent testing led to better performance (Phelps, 2012).

Researchers and educators have speculated on the mechanisms behind the benefit of frequent testing, and we believe the following may be relevant here. Some have suggested that getting more frequent feedback can prompt students to change their studying habits before the final exam (Hattie and Timperley, 2007). The most common benefit cited for increasing test frequency is to help students avoid procrastination in their study efforts (Leeming, 2002; Son, 2004; De Paola and Scoppa, 2011). It could be argued that testing on a weekly basis encourages students to space their studying over cramming. Extensive research on spaced studying suggests that studying the same material repeatedly over a long period of time leads to higher performance over massed studying, especially when the delay between studying and testing is long (for a review, see Son, 2004). Zechmeister and Shaughnessy (1980) suggested that students space their studying based on a metacognitive decision of how well they know the information. Thus, allowing for more frequent exams should theoretically allow for better metacognitive awareness.

If students benefit from frequent, cumulative exams, then it is logical that the benefit will be enhanced by a belief that one can succeed despite early failures and by implementing metacognitive practices that allow one to transform these failures into meaningful learning experiences (Kitchen *et al*., 2003; Dweck, 2006; Duckworth *et al*., 2007; Tanner, 2012; Hochanadel and Finamore, 2015). Unfortunately, a student's willingness to use early failures productively can be undermined by including those early exams in the course grade. If low scores on early assessments make it impossible for a student to reach his or her ideal course grade, the student may get discouraged and stop putting forth the effort required to succeed. Grades often carry the burden of defining a student's self-worth, and the fear of failure can dampen students' intrinsic motivation to learn (Covington and Mueller, 2001). On the other hand, students may not invest sufficiently to benefit from the testing effect if exam scores do not count toward the final grade and the stakes seem too low. Moreover, placing all the grading emphasis on the final summative exam is frightening for many students. These challenges are likely to be magnified at the maturity level of first-year students. We anticipated that a creative course-grading scheme that made the exams high stakes yet allowed for low scores to be dropped would provide the benefits of the testing effect without discouraging the students.

Accordingly, we embarked on a multiyear effort to employ a formative course format in a first-year biology course that already emphasized scientific reasoning skills. The new format converted the two traditional midterm exams into shorter weekly exams with immediate feedback and added creative grading schemes that allowed for early failures without penalty (Figure 1). We hypothesized two impacts of this change in course format:

1. The formative course format will enhance student ability to solve problems that require scientific reasoning skills.
2. The formative course format will not induce negative student attitudes about the course.

Hypothesis 1 was tested by comparing scores on final exam items that were common between the original and formative formats of the course. Hypothesis 2 was addressed using data from student course ratings.

MATERIALS AND METHODS

Course Design

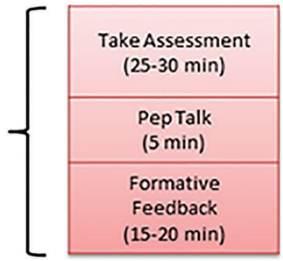
PDBIO 120 is an introductory biology course required for students anticipating a major in various areas of the life sciences and nonmajors with high interest in a premedical or predental curriculum. The course is designed to provide students with a thorough understanding of foundational concepts, including cell theory and biological compartments, the central dogma of molecular biology, biotransformation of energy, reproduction (mitosis, meiosis, etc.), genetics, and evolution. Students are also expected to think critically about scientific studies and master an understanding of some of the defining characteristics of biological experiments, for example, hypotheses, controls, independent and dependent variables, distributions, *p* values, and correlation.

An outline of the topics to be covered during each lecture was provided, and students were required to complete the corresponding reading assignment from the text (Freeman, 2005, 2010) before each class period. Students self-reported completion of these reading assignments to receive credit. To promote a deeper understanding of concepts presented, we required students to participate in "elaborative questioning" for a minimum of 1 hour per week (patterned after the "elaborative interrogation" technique; see Pressley *et al*., 1988; McDaniel and Donnelly, 1996). The exercise was completed outside lecture and allowed each student to actively engage with another individual to explain and ask each other thought-provoking questions on challenging concepts. Students self-reported completion of this activity and turned in a written prompt about the topic to get credit. Students were trained on how to complete these assignments with in-class demonstrations and discussions near the beginning of the semester. In addition to elaborative questioning, students were assigned homework problems designed to provide practice on analytical concepts before midterm exams. Students were responsible for completing the problems, checking their own answers using a provided key, and certifying that they did so before the date specified. Full points were given for completion regardless of their scores. Time was reserved outside the regularly scheduled class periods to provide students with opportunities to complete elaborative questioning and receive help from the instructor and teaching assistants on the

A: Typical Week

Weekend	Mon	Tues	Wed	Thu	Fri
Assigned Reading	In-Class Learning Activities	Assigned Reading	In-Class Learning Activities	Homework Practice	In-Class Assessment

B: Friday Schedule



C: Grading Schemes

Scheme	Attendance	Reading	Homework	Assessments	Final Exam	% of Class
Consistent Performer	5%	5%	10%	40% (all)	40%	0.5%
Improver	5%	5%	10%	30% (top 5)	50%	41.8%
Late Bloomer	10%	10%	10%	0%	70%	33.2%
Self-Teacher	0%	0%	0%	40% (top 5)	60%	20.1%
Extreme Self-Teacher	0%	0%	0%	0%	100%	8.7%

FIGURE 1. Course structure with the new weekly midterm format. (A) Layout of a typical week during the semester showing both in-class (blue text) and out-of-class (black text) learning activities. (B) Details of the class structure on assessment days. "Pep talk" refers to efforts by the instructor to encourage students to reflect and take proactive steps to improve based on the outcome of the assessment. "Formative feedback" includes explaining answers to problems and addressing student questions and concerns. (C) Details of the five grading schemes used in the course. "% of class" refers to the proportion of students for whom that scheme was the most advantageous in a typical semester.

assigned homework problems if necessary. A nongraded (nonrequired) course pretest was available to students at the beginning of the semester. Students were strongly encouraged to take the test, as it consisted of problems drawn from previous exams and offered insight on the level of understanding expected during the course.

Original Format versus Formative Format

Before 2006, PDBIO 120 maintained a traditional exam format, in which two noncumulative, selected-response midterm exams were administered during the semester. Questions on each exam preserved the fundamental quality of assessing students' understanding of the material and their ability to think analytically. This format was replaced by a new structure in which traditional midterm exams were eliminated and supplanted by weekly formative midterms administered during the scheduled class time on Fridays (see Figure 1A).

Each formative midterm exam consisted of 10–20 selected-response problems testing students' proficiency in concepts introduced during the week and included material assessed previously. On average, 60% of the exam questions required higher-level skills according to Bloom's taxonomy, and 40% were low-level questions (Anderson and Krathwohl, 2001). Identical problems were never repeated; rather, new questions were designed to give students practice on the same concepts and skills that had been tested previously. For example, students saw the test item shown in Supplemental Figure S1 on one test, then the items shown in Supplemental Figures S4 and S5 on future exams for more practice. All three items required the same data-analysis skills in order to draw a conclusion, but each included different contexts and different data (and Supplemental Figure S5 required a type of statistical test different from that of the other two examples). While students saw a test item like Supplemental Figure S2 on one exam, future midterms would include similar questions, but the students would be given the sequence of a different molecule in the process (perhaps the nontemplate strand of DNA, the messenger RNA [mRNA] or the anticodon) and would be asked to determine the sequence of another. Furthermore, students could be given sequences at the beginning of a gene instead of a sequence found in the middle, or the sequence could be written with the 3′ end on the left. As one last example, students saw a test item like Supplemental Figure S3 on one midterm, but future exams would give the weight of a different molecule (mRNA or gene) and ask for one of the other molecule's weight. Thus, the test question required the same content knowledge and similar skills, but students could not simply memorize one algorithm in order to succeed every time.

To achieve an exam-like atmosphere for each formative midterm, we gave students 25–30 min to complete the exam. Answer sheets were then collected, the instructor gave a short pep talk designed to increase metacognition and promote a growth mind-set, and the remaining 15–20 minutes were used to provide feedback on each exam item (see Figure 1B). During the feedback portion of class, students could discuss test items with their peers, and then the instructor provided correct answers and explanations.

In these pep talks, the instructor explicitly encouraged students to learn from their mistakes and grow rather than give up. The instructor asked the students to think about what they could do during the coming week to prepare for the next assessment. Students were invited to identify specific test items and concepts that were difficult for them and to seek out help from the instructor or the teaching assistants. The instructors may have shared personal experiences in which they performed poorly on an exam and learned that they needed to study differently in order

to succeed. As another example, the metaphor of baking a cake could be used. If one opens the oven and finds that the cake is not done, the cake is not thrown out. No one criticizes the cake for not being done; rather, the cake just needs to bake longer until it is done. The students are given the message that they should not criticize themselves or assign early judgment about their abilities. They are on the road to mastery and need to just keep working hard.

Learning objectives, reading and homework assignments, and the final summative exam (65 questions, ~60% high-level questions and ~40% low-level questions; Anderson and Krathwohl, 2001) were fundamentally identical to those in the original format of the course. Grades were determined based on student performance relative to a standard rather than in competition with classmates. To reward students for improvement and achievement, we calculated final grades were calculated under five grading schemes, which each assigned different weights to course work, midterms, and the final exam (see Figure 1C). The grading scheme providing the highest overall percentage for each student (as determined with a simple Excel spreadsheet) was used automatically to determine each student's final grade in the course.

Data Collection

Data were collected during three semesters of the weekly formative midterm structure (2011–2013) and two semesters of the previous traditional exam structure (2005). These semesters were chosen because the course pedagogy had become stable following either initial course development or incorporation of the new format. Three instruments were used to compare the overall effectiveness of these two designs: first, quantitative analysis of student performance on common items found on midterms and the final exam (45 items: 18 classified as "remember," 14 "understand," and 13 "apply/evaluate"); second, affective data obtained from university course evaluations collected anonymously at the conclusion of each semester; and third, voluntary individual student interviews offered insight on student metacognition that course evaluations could not provide. Because we were interested in the stories of those students the course format seemed to help, we invited students who were consistently successful throughout the semester (three of these students agreed to be interviewed) and students who improved throughout the semester (five of these students agreed to be interviewed). Thus, these interviews cannot be generalized to the whole class. The interview portion of the study was carried out by an undergraduate research assistant, and names of participants were withheld from instructors to avoid intimidation or desires to please the professor. Interview questions are available in the Supplemental Material.

All error bars represent the SE or the range (when $n = 2$) for the semesters tested. This study was approved by the institutional review board on December 14, 2005, and again on March 7, 2013.

RESULTS AND DISCUSSION

Formative Course Format

To implement the formative format (Figure 1), we incorporated 10 midterm exams, each designed to feel as important to students as traditional graded exams (as opposed to "practice" exercises or quizzes). Because the highest grade from among the five schemes was automatically given, students were rewarded for success, not penalized for failures, and attention was directed toward the final exam (Figure 1C). Immediate detailed feedback was provided in class after each exam to help students resolve misconceptions and make plans to improve. A weekly pep talk was included to offer encouragement and instruction in metacognitive planning (see *Materials and Methods* for details). Finally, each midterm was comprehensive to eliminate the idea that the course was segmented, to provide recurrent practice on difficult concepts, and to help students determine whether previous misconceptions had been fully resolved.

Table 1 summarizes the five semesters included in this study, including class sizes, "D" letter grade rates, fail rates, and correlations between various course requirements and final exam scores. For all semesters, there was approximately one teaching assistant for every 30 students. The rate of "D" letter grades was greater after the formative course format was implemented, but failure rates were not significantly different between course formats. In both cases, rates were very low. Midterm exams scores were the best predictors of final exam performance regardless of course format. Interestingly, homework completion and reading assignment completion were better predictors of final exam performance as part of the formative course format compared with the original course format, even though these activities were identical. Attendance was not recorded during the first semester of the study, but the rest of the data suggest that attendance may have been a better predictor of performance in the formative course format compared with the original course. Even so, homework completion, reading completion, and attendance were likely not the main drivers of increased student performance, since the correlation coefficients were relatively low.

Aggregate Exam Scores

Figure 2A demonstrates that performance on the final exam improved significantly with weekly midterm exams compared with the original format of two midterms. Importantly, the magnitude of improvement depended on the type of item, based on analysis of variance (ANOVA). For items at the lowest Bloom's level (remember; Anderson and Krathwohl, 2001), performance with the weekly midterm format was not statistically better than that with the original format. In contrast, benefits of the new format were observed for items assessing comprehension (understand, 9% gain) and particularly for those requiring application/evaluation (17%, see Supplemental Figures S1–S5 for example test items). Because it is generally accepted that test items using higher levels of Bloom's taxonomy (such as apply and evaluate) require both content knowledge and critical-thinking skills, it appeared the weekly midterm format was especially effective at helping students develop the scientific reasoning skills emphasized in the course (Zoller, 1993; Crowe *et al*., 2008).

Figure 2B shows the temporal basis for this benefit, comparing average scores during the semester for one type of item included in the apply/evaluate category (Figure 2A). Specifically, these problems required students to evaluate experimental data and draw the appropriate conclusion (see Supplemental Figure S1). These results demonstrate that student ability to solve these problems developed gradually and did not reach a plateau until about six iterations. In fact, had we stopped with three iterations (the equivalent of the more traditional two

TABLE 1. Course requirements and learning activities as predictors of student success[a]

Format	Semester	Class size (n)	"D" grade rate (%)	Fail rate (%)	Correlation with final exam score			
					Homework completion	Midterm scores	Reading completion	Attendance
Original	Winter 2005	143	0.7	2.1	$p = 0.02$, $r^2 = 0.04$	$p < 0.0001$, $r^2 = 0.54$	$p = 0.16$, $r^2 = 0.01$	Not recorded
	Fall 2005	148	0.0	0.0	$p = 0.34$, $r^2 = 0.01$	$p < 0.0001$, $r^2 = 0.54$	$p = 0.20$, $r^2 = 0.01$	$p = 0.03$, $r^2 = 0.03$
Formative	Fall 2011	200	4.5	2.0	$p < 0.0001$, $r^2 = 0.10$	$p < 0.0001$, $r^2 = 0.42$	$p < 0.0001$, $r^2 = 0.13$	$p < 0.0001$, $r^2 = 0.14$
	Fall 2012	224	3.1	0.9	$p < 0.0001$, $r^2 = 0.08$	$p < 0.0001$, $r^2 = 0.53$	$p = 0.0003$, $r^2 = 0.06$	$p < 0.0001$, $r^2 = 0.19$
	Fall 2013	165	6.1	0.6	$p < 0.0001$, $r^2 = 0.16$	$p < 0.0001$, $r^2 = 0.67$	$p < 0.0001$, $r^2 = 0.15$	$p < 0.0001$, $r^2 = 0.19$

[a]Linear regression was performed, with final exam score as the dependent variable and various course requirements (homework completion, including elaborative questioning assignments and packet problems; midterm exam scores; completion of reading assignments; and attendance) as independent variables.

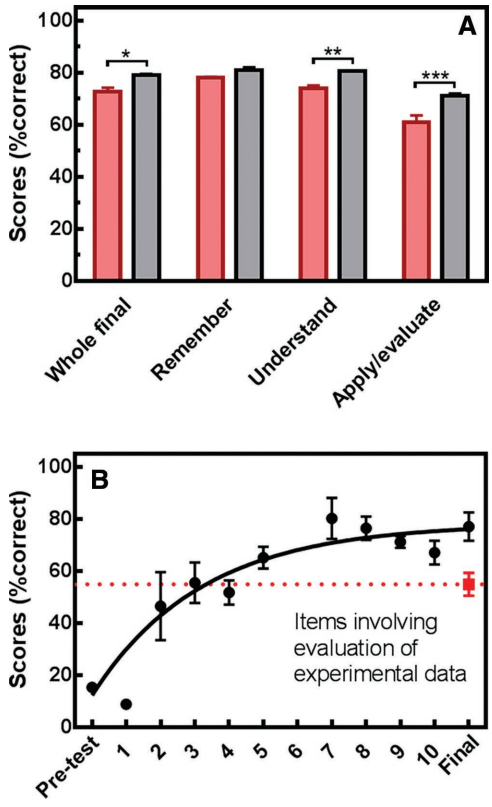


FIGURE 2. Student performance before and after implementation of the new course format. (A) Student performance on final exam items that were common before the intervention (red bars) and after full and stable implementation of the new format (gray bars). The "whole final" difference after the intervention was significant by t test ($n = 2–3$; *$p = 0.015$). The various categories of items (remember, understand, and apply/evaluate) were compared among themselves and between course formats by two-way ANOVA. The main effects of course format and item category were both significant ($p = 0.015$, 22% of the variation; $p < 0.0001$, 72% of the variation) with a small interaction ($p = 0.02$, 5% of the variation). The understand and apply/evaluate categories were individually different between course format based on the Fisher least significant difference (LSD) post hoc test (**, $p = 0.004$; ***, $p = 0.0002$). (B) Student performance on one type of apply/evaluate item ($n = 3$). Red represents performance on the final exam items during the preintervention semesters ($n = 2$).

midterms and a final exam), the performance would have been the same as observed in previous years before the formative midterm format was adopted (red dotted line, square). The incremental increases in student success did not simply reflect low-level memorization of patterns, since the conceptual, experimental, and data-presentation contexts of these items were different in every case on both exams and practice problems (e.g., compare Supplemental Figures S1, S4, and S5). Similar results were observed for simpler apply/evaluate items (see Supplemental Figures S2 and S3 for examples), although the rate of improvement was often greater than that observed in Figure 2B.

Individual Student Performance Trajectories

While the data of Figure 2 show the overall benefits of the weekly midterm format for the class, they do not reveal how

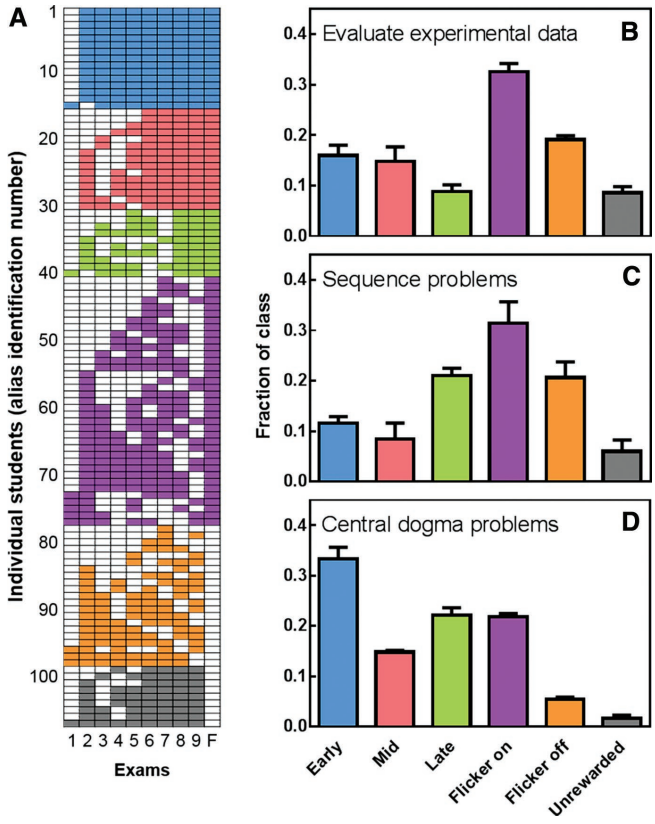


FIGURE 3. Diversity of individual student learning pathways observed with the formative course format. (A) Example heat map showing individual student performance (rows) on the evaluate experimental data items referenced in Figure 2B (sequentially in columns, nine midterms then final exam; white indicates incorrect answer; from 2013; see Supplemental Figures S1, S4, and S5 for example problems). Students were classified into learning categories as follows: blue, early (consistent success on last 80% of exams), pink, mid (consistent success on last 50% of exams), green, late (consistent success on last 30% of exams), gray, unrewarded (would have been early or mid but incorrect on final exam), violet, flicker on (any other pattern, correct on final exam), orange, flicker off (any other pattern, incorrect on final exam). (B) Average fraction of the class included in six categories of learning for evaluate experimental data items ($n = 3$ semesters). Not included were students who did not take all of the midterm exams. (C and D) The analysis of B was applied to sequence problems (C; see Supplemental Figure S2 for example item; one semester only had seven total exams with this type of problem, so in that case late learners were defined as succeeding on the last 40% of exams to ensure true stability) or central dogma problems (D; see Supplemental Figure S3 for example item).

that benefit is realized for different students. We used heat maps to examine and compare performance for each student individually during the semester. A total of nine heat maps were generated: one for each of three different types of exam item ("evaluate experimental data," "sequence problems," and "central dogma problems") for each of three semesters. Figure 3A displays one of those heat maps as an example. In this case, the map details performance on test items involving evaluation of experimental data for each of the students in the 2013 offering of the course. Each row represents a different student in the class. Each column represents a different exam (shown sequentially, left to right, from the first assessment through the final exam). If the student missed the item, the corresponding cell was marked with white. If he or she succeeded, it was marked with color. For example, Student 40 answered the data-evaluation item correctly on exam 1, incorrectly on exam 2, correctly on the next three exams, incorrectly on the next two, and correctly on the last three, including the final exam. Student 10 was successful on the data-evaluation items on each exam after answering incorrectly on the first exam.

The heat maps were then sorted to search for any patterns of student performance on this type of item during the semester. The analysis showed that individual student learning trajectories were diverse and could be distinguished based on six general patterns of learning. The different patterns we were able to identify are marked with different colors in the figure. Students were classified as "early" (blue; consistent success on the last 80% of exams, including the final), "mid" (pink; consistent success on the last 50% of exams after early miscues), or "late" (green; consistent success on the last 30% of exams after early miscues). Students who would have been early or mid learners but failed to succeed on the final exam were classified as "unrewarded" (gray). Those with any other pattern were labeled as "flicker on" (violet; correct on final exam) or "flicker off" (orange, incorrect on final exam). For the data-evaluation problems, the proportions of students in each of these six categories were consistent across three semesters (see small error bars in Figure 3B). Interestingly, the mid and late learners (those who did not show stable mastery until at least halfway through the semester) comprised 24% of the class on average, a level approximately equal to the gain we observed on this type of item on the final exam with the formative midterm scheme compared with the original course format (22%, Figure 2B). These results suggest that the mid and late learners are the students who would not have succeeded on this type of item on the final exam in the original course format. They have therefore benefited specifically from the increased number of chances to iterate. Importantly, this is a quarter of the class.

Figure 3, C and D, compares student learning patterns for other types of apply/evaluate problems. The sequence problems (Figure 3C; see Supplemental Figure S2 for sample) involved inferring DNA, RNA, or protein sequences from each other. The central dogma items (Figure 3D; see Supplemental Figure S3 for example) required calculation of the molecular weight of genes, mRNA, or proteins when the weight of one is known. As demonstrated by the relatively small size of the error bars, the proportions of students in each category were reproducible from semester to semester for a given type of problem. Nevertheless, the distribution of these proportions was unique to each item type. These differences could relate to the complexity of the task. For example, when the problem was more formulaic (Figure 3D), the proportion of early learners was greater. Even so, a large proportion of the class (37%, sum of mid and late) benefited from multiple midterm exams.

It is clear that students who "flicker" (Figure 3, violet and orange) are gaining from the experience; the density of correct answers increases among these students throughout the semester, and more than half of the "flickering" students succeeded on the final. Although this is true for the aggregate, individual

students have no power to predict from their own midterms whether they will be a member of the flicker on cohort or the flicker off cohort, since a stable pattern has not yet been established. However, based on our experience, there is a danger that the individual feels encouraged enough by occasional successes to be lulled into believing that mastery has occurred. In fact, evidence for this idea can be seen in Figure 3A by comparing performance on the last midterm with performance on the final exam for the flickering students. In the flicker off category, 57% of the students answered correctly on the last midterm exam. However, only 30% of those in the flicker on category succeeded on the last midterm. This disparity was reproducible for the different problem types and among the three semesters illustrated in Figure 3 ($p = 0.01$, 49 ± 7% in flicker off and 21 ± 9% in flicker on). We hypothesize that many flickering students who succeeded on the last midterm may have falsely assumed mastery, while many who failed were alerted to the need for more work. Accordingly, students who are flickering in their performance need to be advised by the instructor that, even though they appear to be on a trajectory toward mastery, they have not yet arrived, and ongoing practice is still needed.

Student Attitudes

Having observed gains in student performance, we were anxious to know whether we were also successful at averting negative emotions associated with multiple midterms and emphasis on the final exam. Previously, this course received high marks in student evaluations (course and instructor ratings ~5–10% above the university average). These favorable scores were not altered by adopting the formative midterm scheme (based on *t* tests, unpublished data). Because the intervention involved changes to exam and grading procedures, we were especially interested in knowing how the student attitudes toward those elements were affected. We therefore compared four specific survey questions related to course grades, assessments, and feedback to each other before and after the intervention, as shown in Figure 4. Two-way ANOVA demonstrated that students felt essentially the same about each of these items (no significant differences among questions or interaction) and that the attitudes were generally more favorable with the formative format of the course ($p = 0.03$). Individually, three of the four questions showed significant elevation of affect at the 0.05 level after the intervention, and the fourth ("useful feedback") was significant at the 0.1 level. On the surface it may appear that the positive attitudes are simply a reflection of grades that have been artificially inflated by the flexible grading scheme. However, student evaluations were conducted before grades were issued, and final grades were calculated using a high standard (≥93% required for an "A"). In fact, in the original course format, we had to normalize grades to avoid deflation due to the difficulty of the tasks, and with the new formative scheme, grades could be assigned based on criterion levels without normalizing. Hence, the overall course grade point average before and after adopting the new format remained at about the median of grades for all sections of the course and similar to or below the overall department average. Thus, instead of experiencing adverse impacts, students appeared to believe that the course was more fair and effective in terms of assessment and feedback procedures.

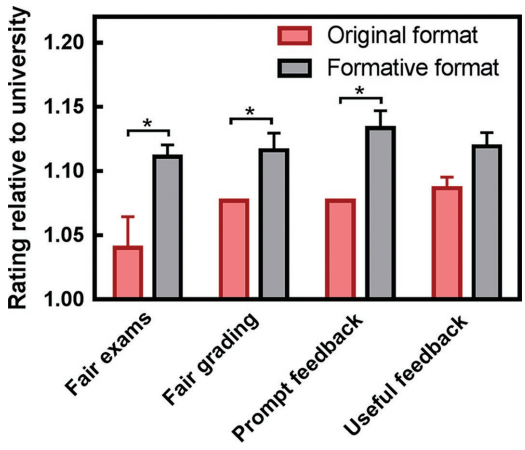


FIGURE 4. Student attitudes toward feedback, exams, and grading improved after implementing the weekly midterm format. Scores from anonymous student course evaluations (standard university rating system) were normalized to scores on the same items obtained across the entire university. The data were analyzed by two-way ANOVA evaluating the main effects of course format and survey question asked. The main effect of course format accounted for 60% of the variation and was significant ($p = 0.03$). No significant difference was found among individual questions; and no format by question interaction was observed. The Fisher LSD post hoc test provided the following *p*-values comparing course format for each question: "fair exams," 0.002; "fair grading," 0.05; "prompt feedback," 0.009; "useful feedback," 0.1. Error bars represent range or SE, $n = 2–3$ semesters; see Figure 2.

We conducted a few (eight) postsemester oral interviews to understand some of the reasons why students might find the new format to be helpful to them (see the Supplemental Material for interview questions). Specifically, students described the advantage of discovering misconceptions without being explicitly penalized (seven of the eight students we interviewed mentioned this: e.g., "Well, it's okay, now I know what I made a mistake on"). All eight students who were interviewed said they wished other classes would adopt a similar format. Interestingly, five of the eight students also expressed a feeling that the format mitigated disappointments and encouraged resilience (e.g., "It wasn't like, 'I bombed this [midterm], so pretty much even if I try I can't get my grade up anymore.'"). All but one of the students mentioned that the course format helped them retain information more than other classes. These interviews provided insights that can inform future research questions.

SUMMARY

Our plan was to use formative assessment and metacognitive emphasis in our introductory course to facilitate student learning. We have shown here that such an intervention can improve exam performance, especially on problems that require higher-order cognitive skills. Moreover, the course rating data suggest that the intervention had a positive effect on student attitudes. As we look toward the future, a number of important questions need to be addressed. For example, several components of the revised course could be individually responsible for the gains in performance, such as the

tests themselves, the unique grading scheme, the immediate feedback, or the weekly pep talks. A critical next step would be to disaggregate these factors by testing them separately. Furthermore, it will be valuable to expand the limited qualitative study to include all class participants one or more years after the course to ascertain the extent to which these practices actually promote and sustain alterations to student self-efficacy and mind-set.

We also learned from this study that performance gains accrue slowly for problems that emphasize scientific reasoning and do not depend on simple recall. Nevertheless, it is clear that the process is highly individual and that certain performance patterns may place students at risk of overestimating their level of mastery. By using this method to examine the entire class, instructors can identify students that could be flickering in their performance and advise them accordingly. Instructors can also show all students the possible learning patterns and explain the risks of flickering, encouraging students to learn from any miscue, even if rare.

Those who would consider adopting this approach may ask, How is this different from the common practice of including weekly quizzes in a course? First, there was a grading scheme that elevated the importance of the exams while allowing for success despite early failures. Second, metacognition was emphasized explicitly and regularly by the instructor. Third, there was immediate feedback and class discussion after the midterm exams. Fourth, the weekly midterm exams were comprehensive. Fifth, there was close alignment between the midterms and the final exam.

Because this formative approach and the associated student-by-exam analysis (Figure 3) are independent of the course content, they can be easily adopted into other courses. Successes in this study and one published previously (Kitchen *et al*., 2003) demonstrate that this approach can be productively adapted for both beginning and advanced students, suggesting that it is amenable to implementation throughout a science curriculum. An advantage to incorporating it into introductory courses is that it may help students alter their approach to exams and learning in a way that will benefit them throughout the remainder of the curriculum.

ACKNOWLEDGMENTS

The contents of this article were developed through U.S. Department of Education (Fund for the Improvement of Postsecondary Education) grant P116B041238. However, the contents do not necessarily represent the policy or opinion of the Department of Education or the government of the United States.

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action. www.visionandchange.org (accessed 22 February 2016).

Anderson LW, Krathwohl DR (2001). A Taxonomy for Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, New York: Addison Wesley Longman.

Balch WR (1998). Practice versus review exams and final exam performance. Teach Psychol 25, 181–185.

Bangertdrowns RL, Kulik JA, Kulik CLC (1991). Effects of frequent classroom testing. J Educ Res 85, 89–99.

Carpenter SK, DeLosh EL (2006). Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. Mem Cognit 34, 268–276.

Carpenter SK, Pashler H (2007). Testing beyond words: using tests to enhance visuospatial map learning. Psychon Bull Rev 14, 474–478.

Carpenter SK, Pashler H, Cepeda NJ (2009). Using tests to enhance 8th grade students' retention of US history facts. Appl Cogn Psychol 23, 760–771.

Carpenter SK, Pashler H, Wixted JT, Vul E (2008). The effects of tests on learning and forgetting. Mem Cognit 36, 438–448.

Carrier M, Pashler H (1992). The influence of retrieval on retention. Mem Cognit 20, 633–642.

Chan JCK, McDermott KB (2007). The testing effect in recognition memory: a dual process account. J Exp Psychol Learn Mem Cognit 33, 431–437.

Covington MV, Mueller KJ (2001). Intrinsic versus extrinsic motivation: an approach/avoidance reformulation. Educ Psychol Rev 13, 157–176.

Crowe A, Dirks C, Wenderoth MP (2008). Biology in Bloom: implementing Bloom's taxonomy to enhance student learning in biology. CBE Life Sci Educ 7, 368–381.

De Paola M, Scoppa V (2011). Frequency of examinations and student achievement in a randomized experiment. Econ Educ Rev 30, 1416–1429.

Duckworth AL, Peterson C, Matthews MD, Kelly DR (2007). Grit: perseverance and passion for long-term goals. J Pers Soc Psychol 92, 1087–1101.

Dweck C (2006). Mindset: The New Psychology of Success, New York: Random House.

Freeman S (2005). Biological Science (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Freeman S (2010). Biological Science (4th ed.). New York: Pearson.

Hattie J, Timperley H (2007). The power of feedback. Rev Educ Res 77, 81–112.

Hochanadel A, Finamore D (2015). Fixed and growth mindset in education and how grit helps students persist in the face of adversity. J Int Educ Res 11, 47–50.

Jensen JL, McDaniel MA, Woodard SM, Kummer TA (2014). Teaching to the test or testing to teach: exams requiring higher order thinking skills encourage greater conceptual understanding. Educ Psychol Rev 26, 307–329.

Johnson CI, Mayer RE (2009). A testing effect with multimedia learning. J Educ Psychol 101, 621–629.

Kang SHK, McDaniel MA, Pashler H (2011). Effects of testing on learning of functions. Psychon Bull Rev 18, 998–1005.

Kitchen E, Bell JD, Reeve S, Sudweeks RR, Bradshaw WS (2003). Teaching cell biology in the large-enrollment classroom: methods to promote analytical thinking and assessment of their effectiveness. Cell Biol Educ 2, 180–194.

Leeming FC (2002). The exam-a-day procedure improves performance in psychology classes. Teach Psychol 29, 210–212.

Marsh EJ, Roediger HL, Bjork RA, Bjork EL (2007). The memorial consequences of multiple-choice testing. Psychon Bull Rev 14, 194–199.

McDaniel MA, Anderson JL, Derbish MH, Morrisette N (2007). Testing the testing effect in the classroom. Eur J Cogn Psychol 19, 494–513.

McDaniel MA, Donnelly CM (1996). Learning with analogy and elaborative interrogation. J Educ Psychol 88, 508–519.

McDaniel MA, Thomas RC, Agarwal PK, McDermott KB, Roediger HL (2013). Quizzing in middle-school science: successful transfer performance on classroom exams. Appl Cogn Psychol 27, 360–372.

Phelps RP (2012). The effects of testing on student achievement, 1910–2010. Int J Test 12, 23.

President's Council of Advisors on Science and Technology (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering and Mathematics, Washington, DC: U.S. Government Office of Science and Technology.

Pressley M, Symons S, McDaniel MA, Snyder BL, Turnure JE (1988). Elaborative interrogation facilitates acquisition of confusing facts. J Educ Psychol 80, 268–278.

Roediger HL, Karpicke JD (2006a). The power of testing memory: basic research and implications for educational practice. Perspect Psychol Sci 1, 181–210.

Roediger HL, Karpicke JD (2006b). Test-enhanced learning taking memory tests improves long-term retention. Psychol Sci 17, 249–255.

Rohrer D, Taylor K, Sholar B (2010). Tests enhance the transfer of learning. J Exp Psychol Learn Mem Cogn Learn Mem Cogn 36, 233–239.

Sedki S (2011). Student preference on exam frequency: a comparative study of St. Mary's University and the American University of Sharjah (UAE). J Int Educ Res 7, 4.

Seymour E, Hewitt NM (1997). Talking about Leaving: Why Undergraduates Leave the Sciences, Boulder, CO: Westview.

Son LK (2004). Spacing one's study: evidence for a metacognitive control strategy. J Exp Psychol Learn Mem Cogn Learn Mem Cogn 30, 601–604.

Tanner KD (2012). Promoting student metacognition. CBE Life Sci Educ 11, 113–120.

Wickline VB, Spektor VG (2011). Practice (rather than graded) quizzes, with answers, may increase introductory psychology exam performance. Teach Psychol 38, 98–101.

Zechmeister EB, Shaughnessy JJ (1980). When you know that you know and when you think that you know but you don't. Bull Psychon Soc 15, 41–44.

Zoller U (1993). Are lecture and learning compatible? Maybe for LOCS—unlikely for HOCS. J Chem Educ 70, 195–197.